AUGUST 1992

ACADEMIC MEDICINE

JOURNAL OF THE ASSOCIATION OF AMERICAN MEDICAL COLLEGES

AAMC

Test of OSCE for Licensure in Canada

A Program in China to Foster Clinical Skills

Using Storytelling to Teach Values and Attitudes

Information and Management System for Students

Community-based Training and Health Services

Professed and "Actual" Teaching Behavior

An Experiential Conception of the GME Curriculum

Computer Analysis of Curriculum Science Content

Natural Science in the Premedical Curriculum

Women Physicians' Partners and Careers

Use of Free Time in a Flexible Curriculum

A Physical Therapist in a Family Practice Residency

VOLUME 67 • NUMBER EIGHT

AUGUST 1992
VOLUME 67
NUMBER EIGHT

ACADEMIC MEDICINE

JOURNAL OF THE
ASSOCIATION OF
AMERICAN MEDICAL
COLLEGES

Special Articles

487 **An Objective Structured Clinical Examination for the Licentiate: Report of the Pilot Project of the Medical Council of Canada**, by RICHARD REZNICK, M.D., M.ED., SYDNEY SMEE, ARTHUR ROTHMAN, ED.D., ANDREW CHALMERS, M.D., DAVID SWANSON, PH.D., LOUIS DUFRESNE, M.D., GUY LACOMBE, M.D., JOHN BAUMBER, M.D., PH.D., PEETER POLDRE, M.D., M.ED., LOUIS LEVASSEUR, M.D., ROBERT COHEN, PH.D., JENNIFER MENDEZ, PH.D., PAUL PATEY, M.D., M.SC., DONALD BOUDREAU, M.D., and MICHEL BÉRARD, M.D.

495 **A New Program to Enhance the Teaching and Assessment of Clinical Skills in the People's Republic of China**, by PAULA L. STILLMAN, M.D., and WILLIAM D. SAWYER, M.D.

500 **Storytelling as a Method for Teaching Values and Attitudes**, by WILLIAM A. HENSEL, M.D., and TERESA L. RASCO

Ideas for Medical Education

505 **Computers to Go**, by ROBERT S. WIGTON, M.D., M.S.

505 **A Computer-based Information Access and Management System for Students**, by LELAND H. PARKS, PH.D., GAROLD O. MINNS, M.D., and ROBERT T. MANNING, M.D.

International Medical Education

507 **The Negev Project in Community-based Training and Health Services: Medical Education Confronts the "Tea Lady,"** by WILLIAM E. SEIDELMAN, M.D., and BASIL PORTER, M.B.B.CH., M.P.H.

Book Reviews

Learning to Play God: The Coming of Age of a Young Doctor, by Robert Marion, rev. by ROBERT G. PETERSDORF, M.D., 511 • **The Ends of Human Life: Medical Ethics in a Liberal Policy**, by Ezekiel J. Emanuel, rev. by NANCY S. JECKER, PH.D., 512 • **A History of Education in Public Health**, ed. by Elizabeth Fee and Roy M. Acheson, rev. by LESTER BRESLOW, M.D., M.P.H., 513

furthermore . . .

514 **Tube Feeding**, by Richard Selzer, commentary by DAVID H. FLOOD, PH.D., and RHONDA L. SORICELLI, M.B., B.S.

From the President

516 **August**, by ROBERT G. PETERSDORF, M.D.

AAMC Focus

517

National Policy Perspectives

518 **Fourth-of-July Fireworks over National Health Care**, transcript of the President's message and response of Senate Majority Leader George Mitchell (D-ME)

Letters to the Editor

A Way to Assure Curricular Quality, by JAMES P. WHALEN, M.D., GERARD CERCHIO, M.D., and HYMAN MUSLIN, M.D., 520 • **A New Problem-solving Curriculum**, by PANAYOTIS G. IATRIDIS, M.D., D.SC., 520 • **A Mentorship Program for Minority Students**, by STEPHEN E. PETERSON, PH.D., and PAUL CARLSON, PH.D., 521

Original Contributions

ARTICLES

522 **What We Say and What We Do: Self-reported Teaching Behavior versus Performances in Written Simulations among Medical School Faculty**, by SANDEE L. HARTMAN and MARC S. NELSON, M.D., M.A.

528 **Conceptualizing Curriculum for Graduate Medical Education**, by JACK ENDE, M.D., and ELAINE ATKINS, PH.D.

535 **Using a Database to Analyze Core Basic Science Content in a Problem-based Curriculum**, by ROBERT L. ROSEN, M.D., HAROLD A. PAUL, M.D., and LARRY J. GOODMAN, M.D.

SHORT PAPERS

539 **The Role of Natural Science in the Premedical Curriculum**, by BRUCE DOBLIN, M.D., M.P.H., and STANLEY KORENMAN, M.D.

542 **Women Physicians in Dual-physician Relationships Compared with Those in Other Dual-career Relationships**, by BONNIE J. TESCH, M.D., JANET OSBORNE, M.D., DEBORAH E. SIMPSON, PH.D., SARA F. MURRAY, M.B.A., and JOANNA SPIRO, ED.D.

544 **Use of Free Time in the Third and Fourth Years of a Flexible Curriculum**, by ROBERT J. MCCOLLISTER, M.D.

547 **Consultative and Educational Roles of a Physical Therapist in a Family Practice Residency**, by LYNN L. HALE, P.T., and CHARLES P. SCHUCH, M.S., P.T.

BRIEF COMMUNICATIONS

Using Imaginary Patients to Test Students' Clinical Documentation Skills, by THOMAS O. STAIR, M.D., and JOHN M. HOWELL, M.D., 550 • **Complementing MEDLINE with Citation Searching**, by MIRANDA LEE PAO, PH.D., 550 • **Osteopathic Interns' Well-being in a Rotating Internship**, by ROBERT A. KIRSLING, M.S., ROBERT L. LITCHFIELD, D.O., and ROBERT E. DRABA, PH.D., 550

Publishing Services: Hanley & Belfus, Inc.
210 South 13th Street, Philadelphia, PA 19107

AUGUST 1992
VOLUME 67
NUMBER EIGHT

ACADEMIC MEDICINE

JOURNAL OF THE
ASSOCIATION OF
AMERICAN MEDICAL
COLLEGES

EDITOR: Addeane S. Caelleigh

DEPUTY EDITOR: Albert Bradford

MANAGING EDITOR: Christopher Greene

STAFF EDITOR: Phil Diamond

EDITORIAL ASSISTANT: Lisa R. Dittrich

PUBLISHER'S EDITOR: Sandra A. Lovegrove

ASSOCIATE EDITORS
Senior Associate Editors
Donald G. Kassebaum, M.D.
Edward J. Stemmler, M.D.
Association of American Medical Colleges

Ideas for Medical Education
Robert S. Wigton, M.D.
University of Nebraska Medical Center

International Medical Education
Arthur Kaufman, M.D.
University of New Mexico School of Medicine

Informatics
Mark Frisse, M.D.
Washington University School of Medicine

Book Reviews
John C. Rose, M.D.
Georgetown University School of Medicine (Retired)

EDITORIAL BOARD
Chairman
Milton Corn, M.D.
National Library of Medicine

Deputy Chair
Nancy Gary, M.D.
Health Care Financing Administration

Philip Anderson, M.D.
University of Missouri-Columbia School of Medicine
Nancy Bennett, Ph.D.
Harvard Medical School
Eta Berner, Ed.D.
University of Alabama School of Medicine
Ruth Bulger, Ph.D.
National Academy of Sciences
Zenaido Camacho, Ph.D.
Baylor College of Medicine
Vincent A. Fulginiti, M.D.
Tulane University School of Medicine
Murray M. Kappelman, M.D.
The University of Maryland School of Medicine
Wayne M. Lerner, Dr.P.H.
The Jewish Hospital at the Washington University
Daniel Masys, M.D.
Lister Hill National Center for Biomedical Communications
Allen H. Neims, M.D., Ph.D.
University of Florida College of Medicine
Herbert Pardes, M.D.
College of Physicians and Surgeons of Columbia University
W. Vickery Stoughton
Duke University Medical Center
Martha Vaughan, M.D.
National Institutes of Health

OFFICERS OF THE ASSOCIATION
Chairman
J. Robert Buchanan, M.D.
Massachusetts General Hospital

Chairman-Elect
Spencer Foreman, M.D.
Montefiore Medical Center

Immediate Past Chairman
William T. Butler, M.D.
Baylor College of Medicine

President
Robert G. Petersdorf, M.D.

Chairman, Council of Deans
Kenneth I. Shine, M.D.
University of California, Los Angeles, UCLA School of Medicine

Chairman, Council of Academic Societies
Kenneth I. Berns, M.D., Ph.D.
Cornell University Medical College

Chairman, Council of Teaching Hospitals
C. Edward Schwartz
University of Nebraska Hospital

Academic Medicine (ISSN 1040-2446) is owned and published monthly by the Association of American Medical Colleges, 2450 N Street, NW, Washington, DC 20037. Publishing services by Hanley & Belfus, Inc., 210 South 13th Street, Philadelphia, Pennsylvania 19107. Second-class postage paid at Washington, DC and additional mailing offices.

Copyright 1992 by the Association of American Medical Colleges. All material subject to this copyright may be photocopied for the noncommercial purpose of scientific or educational advancement. Printed in the U.S.A.

Subscription rates: U.S., Canada, Latin America—$60.00/1 year; $110/2 years; $155/3 years; $30.00/year for students; and $7.00 per single copy. All other locations—$70.00/1 year; ($40.00/1 year for students); $130/2 years; $185/3 years; and $7.00 per single copy.

Subscriptions received before June 30 will be honored for the calendar year: (January through December). Issues from January up to receipt of first issue of new subscription will be mailed automatically.

Subscription requests received after June 30th will be honored for the fiscal year (July through June). Subscribers who want back issues for January through June of that calendar year must purchase them as individual back issues at $7.00 each and as available in stock.

All advertising material published in this journal is intended to conform to ethical standards, but acceptance does not imply endorsement. The views expressed in all articles and features in this journal are those of the authors and do not necessarily reflect the position of the AAMC.

CORRESPONDENCE
Manuscripts: The Editor, *Academic Medicine*, Association of American Medical Colleges, 2450 N Street, NW, Washington, DC 20037.
Advertising: Diane Sherel, Advertising Manager, Hanley & Belfus, Inc., 210 South 13th Street, Philadelphia, PA 19107; (215) 546-0313.
Subscriptions and changes of address: Membership and Publication Orders, AAMC, 2450 N Street, NW, Washington, DC 20037; (202) 828-0416.
Cancellations of an *Academic Medicine* subscription must be submitted in writing to Publication Orders, AAMC, 2450 N Street, NW, Washington, DC 20037. Claims for back issues will be honored for only the three months before the cancellation.
Reprints: Request reprints directly from the authors of articles, not from *Academic Medicine*.
POSTMASTER: Send address changes to *Academic Medicine*, AAMC, 2450 N Street, NW, Washington, DC 20037.

RICHARD REZNICK, M.D., M.Ed., SYDNEY SMEE, ARTHUR ROTHMAN, Ed.D., ANDREW CHALMERS, M.D., DAVID SWANSON, Ph.D., LOUIS DUFRESNE, M.D., GUY LACOMBE, M.D., JOHN BAUMBER, M.D., Ph.D., PEETER POLDRE, M.D., M.Ed., LOUIS LEVASSEUR, M.D., ROBERT COHEN, Ph.D., JENNIFER MENDEZ, Ph.D., PAUL PATEY, M.D., M.Sc., DONALD BOUDREAU, M.D., and MICHEL BÉRARD, M.D.

An Objective Structured Clinical Examination for the Licentiate: Report of the Pilot Project of the Medical Council of Canada

Abstract—The Medical Council of Canada (MCC) administers a qualifying examination for the issuance of a license to practice medicine. To date, this examination does not test the clinical skills of history taking, physical examination, and communication. The MCC is implementing an objective structured clinical examination (OSCE) to test these skills in October 1992. A pilot examination was developed to test the feasibility, reliability, and validity of running a multisite, two-form, four-hour, 20-station OSCE for national licensure. In February 1991, 240 volunteer first- and second-year residents were tested at four sites. The candidates were randomly assigned to one of two forms of the test and one of two sites for two of the four sites. Generalizability analysis revealed that the variance due to form was 0.0 and that due to site was .16 compared with a total variance of 280.86. The reliabilities (inter-station) were .56 and .60 for the two forms. Station total-test score correlations, used to measure station validity, were significant for 38 of the 40 stations used (range .14–.60). The results of the OSCE correlated moderately with the MCC qualifying examination; these correlations were .32 and .35 for the two test forms. Content validity was assessed by postexamination questionnaires given to the physician examiners using a scale of 0 (low) to 10 (high). The physicians' mean ratings were: importance of the stations, 8.1 (SD, 1.8); success of the examination in testing core skills, 8.1 (SD, 1.6); and degree of challenge, 7.8 (SD, 2.1). The results indicate that a full-scale national administration of an OSCE for licensure is feasible using the model developed. Aspects of validity have been established and strategies to augment reliability have been developed. *Acad. Med.* 67(1992):487–494.

Sir William Osler said, "To study the phenomena of disease without books is to sail an uncharted sea, while to study books without patients is not to go to sea at all."[1] A corollary to this is that we need to observe students interacting with patients to assess their clinical skills. A fundamental truth is that the closer an examination simulates the eventual task, the more valuable the assessment will be. The heavy reliance in medical education over the last three decades on multiple-choice examinations has resulted in assessment methods that are highly reliable but lack face validity. That is, they do not closely simulate tasks that prospective physicians are expected to perform.

In the last 15 years, two interdependent techniques, patient simulation and the objective structured clinical examination (OSCE), have emerged. The inherent value of standardized patients (SPs) is the possibility of repeatedly and reliably simulating a patient's problem. The strength of the OSCE is the flexibility of its format and the structure it brings to clinical skills assessment. It is the state-of-the-art technique for assessing clinical competence.[2]

Many educators feel the time is right to incorporate performance-based testing into the licensing process. Currently, four major North American testing bodies, The National Board of Medical Examiners, The Educational Commission on Foreign Medical Graduates, The Medical Council of Canada, and La Corporation Professionnelle des médecins du Québec, either are developing or have implemented programs aimed at incorporating clinical skills assessment into their examinations.

Background

Over the last five years, the Medical Council of Canada (MCC) has been reviewing its examination objectives. It became evident that many of the competencies expected of licentiate candidates were not being evaluated by paper-and-pencil–based tests. This is particularly true of clinical skills such as history taking, physical examination, and communication. To date, the Qualifying Examination of the Medical Council of Canada (MCCQE) consists of four papers, labeled Q1–Q4. The first three are multiple-choice–based tests covering

For more information about the authors, please turn to the end of this report on page 493.

Correspondence and requests for reprints should be addressed to Dr. Reznick, Chair, Part II Test Committee, Medical Council of Canada, Toronto General Division, EN9-237, 200 Elizabeth Street, Toronto, Ontario M5G 2C4, Canada.

a broad spectrum of medical knowledge along traditional disciplinary lines. The fourth part consists of a newly developed format aimed at assessing clinical reasoning skills. In keeping with its mandate to establish and promote a qualification in medicine, the MCC initiated a process in 1989 to add to its examination a fifth component that would assess clinical skills.

The new component would be administered a minimum of 12 months after candidates' graduation. A national committee of physicians and educators was appointed by the MCC to review available testing methods. They selected the OSCE format because it was considered uniquely capable of assessing clinical skills, and because it complemented the other components of the MCCQE that assess knowledge-base and clinical reasoning skills. Its inclusion in the MCCQE would permit a more comprehensive evaluation of physicians' competence.

The OSCE was first described by Harden and colleagues in 1975.[3] This method consists of a series of stations in which a candidate performs discrete clinical tasks while being observed and marked from structured checklists. The tasks center around the interaction of a candidate with an SP. The SP is usually a lay individual trained to portray a patient's problem consistently and repeatedly. Much work has been done to refine the technique of SP portrayals,[4,5] and their fidelity has been established.[6–8] The OSCE has emerged as an excellent method of assessing clinical skills. Although labor-intensive and costly, it has been used at the initial phases of teaching history taking and physical examination,[9] at the end of medical school,[10,11] and during residency.[4,12] The OSCE has been used mostly at individual institutions for the assessment of clinical skills. There have been a few attempts to implement OSCEs across several institutions,[13,14] but they have not been used on a national scale for licensure.

The main challenge facing the planners of this project was to design a feasible, reliable, and valid OSCE for the estimated 2,200 candidates per year at a reasonable cost per candidate. The paper describes the results of the pilot project that was undertaken to evaluate the model proposed for national administration.

Method

The committee reviewed OSCE formats currently being used and from these developed a model. The pilot project was designed to test the feasibility of the model and to address several fundamental research questions.

Development of the Model

The committee designed an OSCE that was four hours long and comprised of two types of stations. There were ten 10-minute patient encounter stations and ten couplet stations. A couplet is a five-minute focused patient encounter followed by a five-minute post-encounter probe consisting of questions about the patient just seen. All stations were based on candidate–SP encounters. Twenty short stations were believed to be ideal,[15] as reliability is a function of both test length and the number of stations in a set time frame.[16] The goal was to have an examination that would reliably test a large number of candidates in the shortest possible time.

The committee realized that the task of assessing 2,200 potential candidates would present logistical problems, including geopolitical issues, examination security, and financial limitations. The examination would be administered at many test centers across Canada, with central coordination. The examination had to be offered in English and French. Equivalent forms of the test were necessary for use on the two test days. Each center would give the examination in two sites, and each site would accommodate 40 candidates per administration. Since the examination would be repeated in the morning and afternoon at both sites, there would be a total of eight administrations with a maximum capacity per center of 320 candidates.

Research Questions

The pilot was designed to answer three main questions: Can the examination be delivered at several sites with no systematic advantage or disadvantage to candidates, no matter which site they used? Can the examination be delivered in two forms with no systematic advantage or disadvantage to candidates, no matter which form of the OSCE they took? Can a four-hour, 20-station OSCE achieve an acceptable level of reliability?

Several additional questions were addressed. Can a criterion-referenced approach be used to set the passing score? Can SPs be used as examiners in some stations? Is there a systematic difference in the way candidates are scored, given that three examiners and three SPs would rotate through each station? What is the relationship between candidates' prior performances on the MCCQE and their performances on the OSCE? Would the station-development process result in examination questions that were realistic, challenging, and important?

Experimental Design

Four examination sites were chosen, two in Toronto and two in Montréal. The two Toronto sites were run independently to simulate sites in different cities. Candidates from Toronto, stratified by school of graduation, were assigned randomly to one of the two sites to test the hypothesis that the examination could be delivered at more than one site with equivalence. To test the equivalence of the two examination forms, the Toronto candidates were randomly assigned to sit the morning or the afternoon administration at which the different forms were used.

The two Montréal sites were administered by one office to evaluate the logistics of testing 160 candidates, on one day, with one form at two locations. This was done to test the proposed 1992 model for site capacity. The Montréal center also focused on whether there was a systematic difference in candidates' scores when

SPs rated compared with when physicians rated. One physician and two SPs rated candidates at each of six stations. One SP rated by observation (as an additional examiner) and the other SP rated by recall after portraying the patient.

Questionnaires to assess the content validity of the stations were distributed to the physician examiners, asking them to rate the stations they had observed according to the importance of each station, its capacity to test core skills, and the degree to which they believed it challenged most candidates.

For two reasons, the committee decided to test 320 candidates for the pilot. First, a power calculation, based on an alpha error of 0.05 and a beta error of 0.1, revealed that a sample size of 100 candidates per center was necessary to detect differences as small as 5%. Second, 160 candidates per center per day was necessary to test the feasibility of the 1992 model. The candidates were paid volunteer first- and second-year residents recruited from the three participating institutions: the University of Toronto, McGill University, and Université de Montréal.

Test Development

The blueprint for the pilot was based on the Medical Council's examination objectives. The objectives had been used to generate a matrix of symptoms and causes that was accessed randomly to generate examination problems for the pilot. The random sample was stratified by age group, in proportions derived from billing data from two provinces. Once an assessment objective was deemed practical for an OSCE format, the committee constructed a rudimentary scenario prescribing the station configuration and the domain of competence to be tested. For example, if the symptom–cause generated was abdominal pain due to appendicitis, the committee might prescribe the following: a young man comes to the emergency room complaining of abdominal pain. It is a couplet station focusing on physical examination and management strategies for such a patient. It was theorized that the best means of achieving equivalence of the two forms of the examination lay in having the author create two stations of similar degrees of difficulty, testing the same domain of competence in the same specialty. For example, an author writing a station on appendicitis would also create a similar station on abdominal pain caused by acute cholecystitis.

Stations (i.e., cases) were written by representatives and their colleagues from the 16 Canadian medical schools. The representatives attended a station-writing workshop facilitated by committee members. In addition to writing stations, all the authors rated each checklist item according to its difficulty and relevance so that the Ebel method[17] of criterion-referencing could be used by the committee to set the passing score. Each station received a three-level review. The first review was by project staff to ensure the feasibility of the station. The second level of review was by two content experts. The third level of review was by a panel consisting of an "OSCE expert," a content expert, a family physician, and another physician. All recommended changes were effected centrally. The stations were then translated into either English or French. To verify the fidelity of the translation process, stations were converted back to the original language by independent translators.

The final examination consisted of stations that tested across all major disciplines, in the competencies of history taking, physical examination, communication, and management. Two to four SPs were trained for each station at each site. There were seven patient trainers engaged in producing the pilot, working in teams of two or three per site. The SPs each received an average of five hours of training. Those SPs who had the additional responsibility of scoring received extra training. Three senior trainers met centrally to review each station and ensure a common understanding of the roles. On the examination day, the physician examiners received a one-hour orientation that was standardized across sites. During this hour, they reviewed their station checklists, conferred with other examiners, and met with the SPs working at the station.

Data Collection and Analysis

Candidates' scores were derived from two sources. First, their performances during the interactions with the SPs were assessed using structured content checklists, global rating scales, and an interpersonal communication scale. Second, post-encounter probes required short, written, free-response answers. These were scored on the day of the examination by physicians, who were provided elaborate rating scales for correct answers, partially correct answers, and incorrect answers. Standardization of responses that were difficult to mark was accomplished by telephone communication across sites.

Answer sheets were optically scanned and statistical analysis was performed.[18,19] Generalizability theory was used to derive reliability and to analyze site and form effects. (Generalizability analysis was performed to deduce norm-referenced coefficients.) Examiner and patient effects were assessed using within-station one-way analysis of variance (ANOVA). The candidates were assigned in random order to rotate through the stations. Therefore, observed examiner and patient effects were attributed to true differences in examiners or patients rather than to differences in the candidates. Repeated-measures ANOVA was used to compare the SPs' scorings with the physicians' scorings. Pearson correlation coefficients were computed to compare each candidate's prior scores on Q1–Q4 with the OSCE scores. Station-total test score correlations were calculated as a measure of station validity.

There were approximately 1.25% of data points missing because some examination items were not completely recorded. Missing data points were replaced by a zero, which was the score assigned to a checklist item in those instances where the candidate

Table 1

Numbers of Personnel and Candidates Participating in the Medical Council of Canada's Pilot OSCE, by Site, 1991

Site*	Standardized Patients	Examiners	Post-encounter Probe Markers	Support Staff	Candidates
Site 1	74	30	4	5	63
Site 2	73	31	4	9	62
Site 3	72	39	3	12	54
Site 4	70	29	5	12	61
Total	289	129	16	38	240

*Site 1 was the Hospital for Sick Children; Site 2 was the Sunnybrook Health Science Centre; Site 3 was the Hôpital St. Luc; and Site 4 was the Royal Victoria Hospital.

had not performed the specific task. Thus, when a datum was missing from a checklist item, it was interpreted to mean that the candidate had not performed the task.

Results

On February 2, 1991, the OSCE was administered at two locations in Toronto: the Hospital for Sick Children and Sunnybrook Health Science Centre; and in Montréal at the Hôpital St. Luc (in French), and at the Royal Victoria Hospital (in English). Table 1 indicates the number of candidates tested and the personnel at each of the four sites. Despite an oversubscription of 15%, designed to compensate for candidates' not being available, there was a shortfall at each site, resulting in a study population of 240 candidates.

Table 2 presents the means and standard deviations of the test scores at each examination location. Table 3 depicts the results of the generalizability analysis for the Toronto candidates, who were randomly allocated to a site and a form. The variance due to site was .16, compared with a total variance of 280.86, indicating that there was virtually no difference attributable to site assignment between the candidates' scores. There was no variance due to form, indicating that candidates were not systematically advantaged or disadvantaged by which of the two examination forms they took.

Table 4 provides the reliability statistics for the various components of the examination. The overall reliability of Form A was .56 and that of Form B was .60.

The cutting scores derived from the Ebel method were 60% for Form A and 59% for Form B. These cutting scores were developed before the examination and without the benefit of actual performance data. The mean scores were 60.46 and 63.06 for Form A and Form B, respectively, indicating that if the cutting scores had been applied, over 40% of candidates would have failed. Before Ebel or any other method is used for criterion-referencing a performance-based examination such as the MCCQE Part II, further work will be required.

There were significant SP effects in 13% of the stations and significant examiner effects in 40% of the stations (alpha = 0.05).

In six stations at the Royal Victoria Hospital site, both the SPs and the physicians rated candidates in an attempt to test the validity of using SPs to score. Three separate scores were obtained for each candidate at these stations, with as many as six forms apiece being completed. This resulted in missing data so that only four stations could be assessed. At two of the four stations, the mean scores given by the physicians were equal to those given by the SP observers; at the other two stations the physicians scored the candidates' performances significantly lower than did the SPs. In three of four stations, the mean scores given by the SP observers were not different from the mean scores given by SPs rating from recall. At one of the four stations, the SPs who rated from recall scored the candidates' performances significantly higher than did the SPs who rated from observation. These six stations were videotaped and these differences can be evaluated in the future.

Several concerns regarding examination validity were addressed. Individual item validation was ascertained using item-total test score correlations. The correlations for the 40 stations of the two forms of the examination ranged from .14 to .51. Twenty-eight of the 40 stations had correlations in the moderate range (above .30). Content validity was in part assessed by questionnaires completed by the physician examiners. Three questions were asked concerning station content, each rated on a scale of 0 (low) to 10 (high). The stations were deemed to be "important," with a mean of 8.1 (SD, 1.8). The sta-

Table 2

Mean Scores of Candidates Who Took the Medical Council of Canada's Pilot OSCE, across Sites and Forms, 1991

Site and Form	Mean Score	(SD)
Hospital for Sick Children		
Form A (English)	61.08	(4.58)
Form B (English)	64.12	(4.10)
Sunnybrook Health Science Centre		
Form A (English)	61.07	(2.85)
Form B (English)	62.00	(4.72)
Hôpital St. Luc		
Form A (French)	60.64	(4.77)
Royal Victoria Hospital		
Form A (English)	59.04	(6.69)

tions were thought to have tested core skills, with a mean of 8.1 (SD, 1.6). The stations were believed to have challenged the candidates, with a mean of 7.8 (SD, 2.1).

The candidates' scores were compared with their performances on the MCCQE that they had taken at the end of medical school. The correlations between the total score on the MCCQE and the OSCE score were .23 for Form A and .26 for Form B; .32 and .35, respectively, when corrected for attenuation, that is, taking into account the reliability of the examinations.

Table 3

Generalizability Analysis of the Effects of Form and Site for 120 Randomly Chosen Candidates at the Toronto Testing Sites, 1991*

Effect	Degrees of Freedom	Variance
Form	1	0.0
Site	1	.16
Candidate: (form × site)	116	8.71
Station: form	38	78.53
Form × site	1	0.0
Site × station: form	38	26.73
(Station: form) × (candidate: form × site)	2,204	166.73
Total variance		280.86

*The variance due to site was .16, compared with a total variance of 280.86, indicating that there was virtually no difference attributable to site assignment between the candidates' scores.

Discussion

The OSCE has emerged as the state-of-the-art method of testing clinical skills. It has the major advantage of observation of actual performance. The use of SPs has enabled this examination to simulate real-world conditions with a high degree of fidelity.

The OSCE has limitations. It is a labor-intensive effort that is costly to run. To implement an OSCE requires a dedicated and knowledgeable team. In general, a lengthy examination of six hours or more is required to achieve high levels of reliability.[2]

In the pilot program, the logistics of staging a multisite OSCE were challenging. The examination was centrally coordinated with peripheral administration. This was an effective administrative model for station development, SP trainer training, and standardization of examination delivery. Decisions were made centrally on an ongoing basis, including resolving last-minute difficulties. However, the scope of the project, the several research agendas, and the restricted time line seemed to dictate a need to distribute several production tasks to different locations. This resulted in inefficiencies and overlapping activities.

Two very important lessons were learned. First, production of this type of examination cannot be rushed. Adequate time is needed for case development, standard setting, translation, and the technical formatting necessary for computer scoring. Our estimate is that production must be started 14–16 months prior to administration. Second, an examination of this scale requires that all aspects of development, production, and site supervision be centralized. The pilot suffered because production tasks, site activities, and data analysis were distributed to too many locations for the central committee to provide adequate support and direction.

The costs incurred during the pilot are not a clear indication of the costs of actually running the examination. Excluding the one-time research and development costs, the current projection of the cost per candidate is $800–$1,200 Canadian ($680–$1,020 U.S.) The exact figure is dependent on the total number of candidates tested, the number of sites used, and the degree of medical school support received.

This cost is higher than others have suggested[20] but is inclusive. The projected cost covers fees paid to station authors and examiners, administrative meeting costs, staff salaries, and related expenses. Such expenses are rarely included in estimating cost per candidate because often faculty have volunteered time, and administrative expenses are absorbed through institutional budgets. It is very important to note that the MCC model could not

Table 4

Reliability Data for Form A and Form B, Overall and for Components, as Used in the Medical Council of Canada's Pilot OSCE, 1991

No. of Units	Components	α* Form A, Sites 1–4	α* Form B, Sites 1 and 2
10	10-minute stations	.39	.34
10	Couplet stations	.45	.59
14	History	.45	.50
6	Physical exam.	.43	.39
10	Management strategies	.31	.58
12	Physician examiner	.45	.44
8	Standardized patient examiner	.38	.51
20	Overall	.56	.60

*Reliability, symbolized here by alpha, is the generalizability coefficient calculated as the variance due to candidates divided by the sum of the variances due to candidates, examiners, and candidate - examiner interaction.

have worked at all without the cooperation of the three medical schools involved, as they contributed space and other necessary resources. The MCC is building a partnership with medical schools so that the examination can be delivered at university-supported sites.

It would seem clear that examination materials must be secure for a high-stakes test. However, an OSCE necessitates that case materials be distributed among SPs and SP trainers; thus, the materials are widely circulated well before the examination date. For the pilot, the concern was in part mitigated by the fact that few individuals had more than one of 20 examination questions. In addition, the SPs were selected with consideration for their understanding of the confidential nature of the materials and were required to sign a document agreeing to respect security policies. While security remains a concern, there is recent evidence to suggest that prior knowledge of the topics covered in the test may not affect results. Colliver and colleagues administered the same examination to their students over a three-week period and found that the students who took the test in the early period did no differently from the students who took it at the end.[21]

The results of this pilot showed no difference in students' scores as a function of examination site. This finding is significant for a national-scale examination and is congruent with two earlier multisite studies,[13,14] although these studies could not randomly allocate students to sites. The equivalence between the present study's sites may be attributed to several factors. Central administration ensured strong similarities in the setups of the sites and in standardizing the content and order of the examination. The senior trainers were experienced in selecting and training SPs, which may be a significant factor in producing reliable SP portrayals.[22] While some investigators have cautioned that differences in SP portrayal may become a problem in a multisite OSCE,[23] the present study showed uniformity in SP portrayal in 87% of the stations. Of the 13% of stations where an SP effect was noted, the majority were ones in which SPs functioned as both patient and examiner. In the future, only physician examiners will be used. Strategies for continued equivalence of sites will include an upgraded and strong central administration, improved resources for SP trainers, and videotapes for examination-day orientation of participants.

The proposed model for national administration requires two separate days of testing to accommodate all the candidates. In order to administer a secure examination two distinct forms are required. Our results indicate that the attempt to create parallel and equal test forms was successful. This supports the practice of having matching cases written by the same author and the careful matching of the forms. Almost all cases were matched for gender, age, domain of medical knowledge, and clinical skill tested.

In the Canadian context, equivalence of examination delivery in French and English is essential. We originally attempted to choose bilingual students at random to take the examination in either of the two languages, but were thwarted by the insufficiency of the number of individuals who considered themselves functionally bilingual. However, the mean score of the English-speaking candidates who took form A of the OSCE was 61.04, and the mean score of the French-speaking candidates who took the same form was 60.64. These nearly identical mean scores provide inferential evidence that language was not a factor in the examination.

Reliability is a fundamental consideration of any assessment process. While there is no strict value that constitutes a reliable examination, there is an emerging consensus that reliability of .80 or greater is desirable for a high-stakes examination. The reliabilities of the pilot examination were .56 and .60 for the two test forms. This falls short of the desired figure but is consistent with most other four-hour OSCEs.[2] Factors known to influence the reliability of this type of examination include total testing time, number of stations, integrity of data collection and processing, consistency of the SPs' portrayals, heterogeneity of candidates' competencies, and consistency of examiners' scorings.

More than one examiner was used at each station, since examiners were asked to work a full day and it was anticipated that too heavy a load would fatigue them, which would affect their performance. The analysis revealed significant interexaminer inconsistencies in 16 of 40 stations. Examiners polled after the test uniformly stated that they would have favored working a half day with no breaks. This would eliminate the need for more than one examiner per station, a strategy that will be adopted for the 1992 administration. Examiners will also receive a more comprehensive orientation to their tasks in order to minimize examiner effects across sites.

Two other aspects of the pilot might have had negative impacts on reliability. Reliability was artificially deflated because the candidates were given zero when data were missing. For future administrations, only one form per station will be used, to minimize missing data. Second, the candidates were paid volunteers and likely to have been a more homogeneous group than the candidates who will take the eventual examination; this may have had an effect, because the more homogeneous the group, the harder it is to achieve high levels of reliability.

Table 4 shows that the reliability of the couplet stations was higher in both forms of the examination than was the reliability of the 10-minute stations. A model that might augment reliability is one in which only the couplet-station format was used. This would eliminate, however, the possibility of assessing competencies such as communication skills and management skills, since these require the longer 10-minute patient-encounter format.

The major consequence of suboptimal reliability is an increased inci-

dence of false positives; that is, an increased possibility that some candidates who were not competent might pass the examination by chance if given enough opportunities to repeat. One way to avoid this is to strictly limit the number of times the examination can be taken. Other options being considered to augment reliability include combining the Q5 score with other elements of the MCCQE for purposes of making a pass-fail decision; and lengthening the examination.[2,16,17]

The concerns about reliability for a high-stakes examination cannot be taken lightly. Although it is anticipated that changes being made in the delivery of the examination will result in a higher reliability than was achieved in the pilot, this is not a certainty. It may be that reliability levels in the range of .60 are an inherent feature of a four-hour OSCE given to a relatively homogeneous population. Despite this concern, it would seem reasonable to include the OSCE in a program of national licensure as part of a comprehensive examination system that assesses different competencies with different examination methods. The OSCE is one of the few examination methods that can assess data-gathering skills and patient-physician communication skills, and is the only method that can achieve this with a high degree of fidelity to the clinical reality.

Validity has many aspects. The OSCE was chosen partly for its high face validity. Examiners confirmed content validity by judging the stations to be important, challenging, and suitable for the testing of core knowledge. Construct validity and criterion-referenced validity were not addressed.

A major unanswered question in performance-based testing is how to set the passing standard. For the last several decades, cutting scores have been norm referenced. This technique sets the passing score with reference to a defined group, resulting in a fixed number of individuals' passing the examination regardless of the difficulty of that particular administration. Within the MCC, there has been growing support for moving away from norm referencing, as it is argued that the passing score set is arbitrary.

The MCC, along with other licensing and certification bodies, is attempting to use criterion referencing. This involves linking success or failure to a fixed standard, usually defined by a panel of experts who attempt to numerically define the minimally acceptable performance. The difficulty in applying criterion referencing to this type of examination centers around two problems. First, there are no prior performance data on which to anchor judgments. Second, judges have intrinsic difficulty, especially in the field of medicine, with the notion of the minimally competent student or practitioner. This difficulty may cause the judges to set unrealistic passing standards, as occurred with the pilot reported here. Criterion referencing makes intuitive sense, however, and offers an approach that is fundamentally more sound than norm referencing. However, its application to performance-based testing is not yet sophisticated enough to be employed with confidence.

Because of difficulties encountered using the Ebel method for setting a passing standard, the MCC is investigating the use of alternative methods of criterion referencing for future administrations of the OSCE. Safeguards are being built in to the decision-making process to ensure that the passing standard set is within acceptable limits.

Clinical skills, such as taking a history, performing a physical examination, communicating with patients and their families, and interpreting data in the context of clinical reality, are central to teaching the art and science of medicine. Licensing authorities have focused on assessing a candidate's cognitive skills, an important but by no means exclusive part of being a good doctor. The results of the pilot affirmed that the model developed for the addition of a clinical-skills-assessment component to the qualifying examination is feasible. The strategies developed for subsequent administrations have the potential for augmenting reliability. There are no plans, however, to run further pilot testing before national administration of the OSCE. The MCC has adopted a resolution establishing the OSCE as the fifth component of a comprehensive examination for the MCCQE, with the first administration in October 1992.

The low-to-moderate correlation between the candidates' scores on this examination and their scores on the multiple-choice component of the MCCQE was provocative. One interpretation is that this new examination is testing other important skills that cannot be inferred from the results on multiple-choice tests. This would support our hypothesis that the OSCE is essential to the licensing process, because it is anchored to the fundamental skills we teach our students.

Dr. Reznick is chair, Part II Test Committee, and assistant professor, Department of Surgery, University of Toronto (UT), Ontario; *Ms. Smee* is national patient training coordinator, Medical Council of Canada (MCC), Ottawa; *Dr. Rothman* is professor, Department of Medicine, UT; *Dr. Chalmers* is head, Division of Rheumatology, University of British Columbia, Vancouver; *Dr. Swanson* is senior evaluation officer, National Board of Medical Examiners, Philadelphia, Pennsylvania; *Dr. Dufresne* is professor of medicine, Université de Montréal, Quebec; *Dr. Lacombe* is associate professor, Internal Medicine-Geriatrics, Université de Sherbrooke, Quebec; *Dr. Baumber* is professor, Office of Medical Education, University of Calgary, Alberta; *Dr. Poldre* is assistant professor, Department of Medicine, UT; *Dr. Levasseur* is professor of surgery, Faculté de Medécin, Université Laval, Quebec City, Quebec, and chair, Central Examination Committee, MCC; *Dr. Cohen* is associate professor, Department of Surgery, UT; *Dr. Mendez* is project administrator, MCC; *Dr. Patey* is professor, Discipline of Family Practice, Faculty of Medicine, Memorial University of Newfoundland, St. John's; *Dr. Boudreau* is assistant professor and associate program director of respiratory medicine, McGill University, Montréal, Quebec; and *Dr. Bérard* is registrar, MCC.

The authors gratefully acknowledge the editorial assistance of Deborah Birkett.

References

1. Bennett, Bean R., and Bennett, Bean W. Sir William Osler: Aphorisms from His Bedside Teachings and Writings. New York: H. Schuman, 1950, p. 76.
2. Van der Vleuten, C. P. M., and Swanson,

D. B. Assessment of Clinical Skills with Standardized Patients: State of the Art. *Teach. Learn. Med.* **2**(1990):58–76.
3. Harden, R., Stevenson, M., Downie, W., and Wilson, G. Assessment of Clinical Competence Using Objective Structured Examinations. *Br. Med. J.* **1**(1975): 447–451.
4. Barrows, H. S., Williams, R. G., and Moy, R. H. A Comprehensive Performance-Based Assessment of Fourth-Year Students' Clinical Skills. *J. Med. Educ.* **62**(1987):80–90.
5. Stillman, P. L., et al. Assessing Clinical Skills of Residents with Standardized Patients. *Ann. Intern. Med.* **105**(1986): 762–771.
6. Norman, G. R., Tugwell, P., and Feightner, J. W. A Comparison of Resident Performance on Real and Simulated Patients. *J. Med. Educ.* **57**(1982):708–715.
7. Vu, N. V., Steward, D. E., and Marcy, M. An Assessment of the Consistency and Accuracy of Standardized Patients' Simulations. *J. Med. Educ.* **62**(1987):1000–1002.
8. Colliver, J. A., Steward, D. E., Markwell, S. J., and Marcy, M. L. Effect of Repeated Simulations by Standardized Patients on Intercase Reliability. *Teach. Learn. Med.* **3**(1991):15–19.
9. Stillman, P., Ruggill, J., Rutala, E., and Sabers, D. Patient Instructors as Teachers and Evaluators. *J. Med. Educ.* **55**(1980): 186–193.
10. Newble, D., Hoare, A., and Elmslie, R. The Validity and Reliability of a New Examination of the Clinical Competence of Medical Students. *Med. Educ.* **17**(1981):165–171.
11. Stillman, P. L., et al. An Assessment of the Clinical Skills of Fourth-Year Students at Four New England Medical Schools. *Acad. Med.* **65**(1990):320–326.
12. Cohen, R., Reznick, R. K., Rothman, A. I., Taylor, B., and Provan, J. Reliability and Validity of the Objective Structured Clinical Examination (OSCE) in Assessing Surgical Residents. *Am. J. Surg.* **160** (1990):302–305.
13. Klass, D., Hazzard, T., Kopelow, M., Tamblyn, R., and Barrows, H. Portability of a Multiple Station, Performance Based Assessment of Clinical Competence. In *Further Developments in Assessing Clinical Competence*; I. R. Hart and R. M. Harden, eds., pp. 434–439. Montréal, Québec: Can-Heal Publications, 1987.
14. Petrusa, E. R., et al. A Multi-Institutional Trial of an Objective Structured Clinical Examination. *Teach. Learn. Med.* **3**(1991): 86–94.
15. Cohen, R., et al. A Comprehensive Assessment of Graduates of Foreign Medical Schools. *Ann. R. Coll. Physicians Surg. Canada* **21**(1988):505–509.
16. Swanson, D. B., and Norcini, J. J. Factors Influencing Reproducibility of Tests Using Standardized Patients. *Teach. Learn. Med.* **1**(1989):158–166.
17. Ebel, R. L. *Essentials of Educational Measurement*. Englewood Cliffs, New Jersey: Prentice Hall, 1972, pp. 492–94.
18. *SPSS/PC PLUS*. Chicago, Illinois: SPSS, Incorporated, 1986.
19. Crick, J. E., and Brennan, R. L. *The Manual for GENOVA*. Iowa City, Iowa: American College Testing Program, 1983.
20. Williams, R. G., et al. Direct, Standardized Assessment of Clinical Competence. *Med. Educ.* **21**(1987):482–489.
21. Colliver, J. A., et al. Test Security in Examinations That Use Standardized-patient Cases at One Medical School. *Acad. Med.* **66**(1991):279–282.
22. Tamblyn, R. N., Klass, D. K., Schnabl, G. K., and Kopelow, M. L. Factors Associated with the Accuracy of Standardized Patient Presentation. In *Research in Medical Education. Proceedings of the Twenty-Ninth Annual Conference*, M. Brownell Anderson, editor-coordinator, pp. S55–S56. Washington, D.C.: Association of American Medical Colleges, 1990.
23. Stillman, P., et al. Assessment of Clinical Skills of Residents Utilizing Standard Patients. A Follow-up Study and Recommendations for Application. *Ann. Intern. Med.* **114**(1991):393–401.

Women Physicians in Dual-physician Relationships Compared with Those in Other Dual-career Relationships

BONNIE J. TESCH, M.D., JANET OSBORNE, M.D., DEBORAH E. SIMPSON, Ph.D., SARA F. MURRAY, M.B.A., and JOANNA SPIRO, Ed.D.

Abstract—This study compared the career and domestic responsibilities of women physicians whose domestic partners were physicians (WP-Ps) with those of women physicians whose domestic partners were not physicians (WP-NPs). In 1988 the authors surveyed 602 women physicians in a large midwestern city regarding their career and domestic roles; 390 were physicians in training (students and residents), and 212 were physicians in practice (academic medicine and private practice). Overall, 382 (63%) responded; of the 382, 247 (65%) had domestic partners; of these 247, 91 (37%) were WP-Ps and 156 (63%) were WP-NPs. The WP-Ps were found to be twice as likely as the WP-NPs to interrupt their careers to accommodate their partners' careers. The WP-Ps also assumed significantly more domestic responsibilities and worked fewer hours practicing medicine than did the WP-NPs. The 163 women physicians in training (44—48%—of the WP-Ps and 119—76%—of the WP-NPs) demonstrated a more egalitarian division of labor overall, with no significant differences between the WP-Ps and the WP-NPs. The authors recommend that longitudinal studies be undertaken to determine whether women physicians in training continue this trend as they enter the practice of medicine. *Acad. Med.* 67(1992):542–544.

The recent expansion of the nation's supply of physicians has brought with it dramatic increases in the number of women entering medical school and practice. By the year 2000, one physician in five will be a woman.[1] A number of studies have examined how women's entry into medicine will affect the delivery of medical care, with particular emphasis on the effects of career interruptions due to childbearing and child-rearing.[2–5] In one study, between 58% and 78% of women physicians were found to have experienced a conflict between marriage or parenthood and career.[4] In two studies, 40–50% of women physicians reported changes in their career plans because of marriage or family responsibilities.[2,3] Another study showed that for women physicians, marriage and child-rearing were strongly associated with reductions in hours worked and with lower earnings.[6]

Approximately 85% of women physicians marry professionals, 50% of whom are also physicians. A limited number of studies have examined dual-physician partnerships.[6–7] According to the 1981 census data, in 73% of these marriages the male physician earns more, and in 63% of the cases, the male physician works more hours. Lorber reexamined these findings and reported an age effect: in couples where both partners were over 40 years old, women physicians tended to follow the lead of men physicians, whereas in couples where both partners were under 40 years old, decisions were more mutual.[7]

The studies just described have outlined the impacts of dual-career relationships on women physicians' career plans and domestic responsibilities. However, there is no evidence that dual-physician relationships have a differential impact on women physicians' careers and domestic responsibilities when compared with other dual-career relationships. The purpose of this study was to compare the career and domestic responsibilities of women physicians whose partners are physicians (WP-Ps) with those of women physicians whose partners are not physicians (WP-NPs).

Method

In 1988 a 28-item questionnaire was developed to elicit information from women physicians in training (medical students and residents) and women physicians in practice (private practice and academic medicine) regarding their career and domestic responsibilities. The questions addressed relationship status and issues, responsibilities for child-rearing and household tasks, and the relative priorities of career, relationship, and child-rearing.

Overall, 602 questionnaires were mailed: 390 to physicians in training affiliated with a large private school in the upper Midwest, and 212 to practicing physicians residing in an upper midwestern metropolitan area (population in excess of 1.2 million). The questionnaire data were keypunched, and measures of frequency, central tendency, and *t*-tests for significance were obtained.

Results

The overall response rate was 63% (382 of the 602 women physicians). Responses by group were as follows: women physicians in training, 57% (222 of 390), and women physicians in practice, 76% (160 of 212). The mean age of the responding physicians was 31 years, with a range of

Dr. Tesch is assistant professor, Department of Medicine; Dr. Simpson is associate professor, Department of Family and Community Medicine; Ms. Murray is research coordinator, Division of Educational Services; and Dr. Spiro is associate professor, Department of Psychiatry and Mental Health Sciences; all at the Medical College of Wisconsin, Milwaukee. At the time of this study, Dr. Osborne was a student at the Medical College of Wisconsin; she is now a resident, Department of Obstetrics and Gynecology, University of Minnesota Hospital and Clinic, Minneapolis.

Correspondence and requests for reprints should be addressed to Dr. Tesch, 8914 Stickney Ave., Wauwatosa, WI 53226.

21–68 years. Of the 382 physicians, 247 (65%) were in marital or other committed relationships; of these 247, 91 (37%) had partners who were physicians, and 156 (63%) had partners who were not physicians. Of the 247 physicians who had partners, 163 were physicians in training, and 84 were physicians in practice. Of the 163 physicians in training who had partners, 44 (27%) had partners who were physicians; of the 84 physicians in practice who had partners, 47 (56%) had partners who were physicians. Thus, the physicians in practice were twice as likely as were the physicians in training to have partners who were physicians.

The results indicate that the 91 women physicians whose partners were physicians (WP-Ps) were more than twice as likely to interrupt their own careers for their partners' careers compared with the 156 women physicians whose partners were not physicians (WP-NPs): 25% versus 11%. In contrast, the men partners who were not physicians were significantly more likely to interrupt their careers than were the men partners who were physicians (41% versus 15%, $p < .05$).

Twelve questionnaire items focused on domestic (e.g., housework) and child-rearing responsibilities (e.g., bathing children, feeding children). On seven of the 12 items, the WP-Ps had significantly more responsibility for childbearing and domestic tasks relative to the WP-NPs (see Table 1).

Seventy percent or more of the WP-Ps had primary responsibility for making child care arrangements, transporting children to the doctor, shopping, and cooking. In contrast, the highest percentage of the WP-NPs who assumed primary responsibility for any single task was 66%, with the lowest percentage being 40%. The range for the WP-Ps was 61–81%. When, among the WP-Ps, the responses of the physicians in training were compared with those of the practicing physicians, the division of labor for various child-rearing and other domestic responsibilities was more egalitarian for the WP-Ps who were physicians in training. Significant differences ($p < .05$) were obtained on nine of the 12 items, with the division of labor for the WP-Ps who were physicians in training closely matching the division of labor reported by the WP-NPs generally.

Table 1

Percentages of 247 Women Physicians Assuming Primary Responsibility for Domestic and Child-rearing Tasks, Comparing Those Whose Domestic Partners Were Physicians with Those Whose Partners Were Not Physicians, 1988*

Domestic or Child-rearing Task	% Assuming Primary Responsibility: Partnered with Physician ($n = 91$)	% Assuming Primary Responsibility: Partnered with Non-physician ($n = 156$)	Difference in % Points
Make child care arrangements	81	66	15†
Transport the children to doctor	76	60	16†
Shop	73	57	16‡
Cook	70	62	8†
Laundry	67	58	9†
Housework	66	58	8†
Transport children to lessons	66	40	26†
Take children to sitter	65	55	10
Go to parent–teacher conferences	64	46	18
Bathe the children	63	56	7
Feed the children	62	55	7
Put the children to bed	61	55	6

*In 1988 the authors compared the career and domestic responsibilities of women physicians whose domestic partners were physicians with those of women physicians whose partners were not physicians. Of the 602 women physicians who were surveyed (consisting of medical students, residents, private practitioners, and academic physicians, all in a large midwestern metropolitan area), 382 (63%) responded; of these, 247 (65%) had domestic partners; of these 247, 91 (37%) had partners who were physicians.

†$p < .05$.

‡$p < .01$.

The increased responsibility in child-rearing and other domestic tasks for the WP-Ps who were in practice was offset by a decrease in hours worked when compared with the hours worked by the WP-NPs. The practicing WP-Ps worked in medicine an average of nine hours fewer per week than did the WP-NPs (43 versus 52). As a corollary, the physician partners of the WP-Ps worked an average of 15 hours more per week than did the non-physician partners of the WP-NPs (58 versus 43).

Discussion

When the women physicians are examined in aggregate, the major findings of this study are consistent with those of previous investigations. However, as just described, when the responses of the physicians are analyzed according to whether their partners were physicians or were not physicians, several important differences emerge.

Perhaps most interesting is the finding that although the WP-Ps generally experienced a significantly greater share in the division of labor for various child-rearing and other domestic responsibilities than did the WP-NPs, those WP-Ps who were physicians in training appeared to enjoy a more egalitarian division of labor — approaching that of the WP-NPs — than did the WP-Ps who were practicing physicians. Longitudinal studies should be undertaken to assess whether women physicians in training whose domestic partners are physicians will maintain a fairly egalitarian division of labor once they become practicing physicians.

References

1. Bowman, M., and Gross, M. L. Overview of Research on Women in Medicine—Issues for Public Policy Makers. *Public Health Rep.* **101**(1986):513-521.
2. Williams, P. A. Women in Medicine: Some Themes and Variations. *J. Med. Educ.* **46**(1971):584-591.
3. Nadelson, C. C., Notman, M. T., and Lowenstein, P. The Practice Patterns, Lifestyles, and Stresses of Women and Men Entering Medicine: A Followup Study of Harvard Medical School of Graduates from 1967 to 1977. *JAMWA* **34**(1979):400-406.
4. Levinson, W., Tolle, S. W., and Lewis, C. Women in Academic Medicine: Combining Career and Family. *N. Engl. J. Med.* **321**(1989):1511-1517.
5. Woodward, C. A., Cohen, M. L., and Ferrier, B. M. Career Interruptions and Hours Practiced: Comparison between Young Men and Women Physicians. *Can. J. Public Health* **81**(1990):16-20.
6. Uhlenberg, P., and Cooney, T. M. Male and Female Physicians: Family and Career Comparisons. *Soc. Sci. Med.* **30**(1990): 373-378.
7. Lorber, J. How Physician Spouses Influence Each Other's Careers. *JAMWA* **37**(1982): 21-26.

Use of Free Time in the Third and Fourth Years of a Flexible Curriculum

ROBERT J. McCOLLISTER, M.D.

Abstract—In 1988, the author surveyed senior students on their use of the six months of free time in the junior-senior biennium at the University of Minnesota Medical School-Minneapolis. The students were asked both to provide detailed reports on their free-time activities and to rank the three activities they considered to be most helpful to their professional development. Of 274 seniors, 191 (70%) responded. The students had used a majority of their free time for pursuits related to education and advancement of their careers. Approximately one-third of the time had been used for vacations. The activities the students viewed as most helpful included research, interviewing for residency positions, vacation, study elsewhere, and preparation for the National Board of Medical Examiners Part II examination. The author suggests that when free time is made available in a flexible curriculum, students use it to advantage in an educationally responsible manner. *Acad. Med.* 67(1992): 544-546.

Within the two full academic years of the junior-senior biennium at the University of Minnesota Medical School-Minneapolis, students are required to complete a total of 72 weeks in full-time assignments in clinical courses. The remaining 24 weeks in the two-year calendar are termed "free time." The inclusion in the M.D. curriculum at this school of what amounts to a half year of free time for students in the junior-senor biennium provides substantially more free-time than is provided in the junior and senior years at most U.S. medical schools. Free time may be used for any purpose, including vacation, and may be scheduled in any manner, from an entire six-month single block to three- or six-week blocks placed throughout the two years. Thus, flexible scheduling at Minnesota provides every opportunity for students in their third and fourth years to schedule their courses and free time to maximum advantage for academic career purposes or personal reasons.

A study of students' use of free time was undertaken to determine how students use the time, with special attention to whether a substantial amount of free time is used for educational purposes to help students progress toward career goals. If so, such flexibility might provide advantages that would recommend this arrangement to other medical schools.

Dr. McCollister is associate dean for curriculum affairs and associate professor, Department of Medicine, University of Minnesota Medical School-Minneapolis.

Correspondence and requests for reprints should be addressed to Dr. McCollister, Box 33 UMHC, University of Minnesota Medical School, Minneapolis, MN 55455.

Method

In April 1988, the 274 senior medical students at the University of Minnesota were surveyed by questionnaire and were asked to provide a detailed analysis of how each had used the 24 weeks of free time. To promote accurate recall, the students were given a copy of the two-year academic program that showed where their free time had been scheduled. Several possible categories of activities were suggested to prompt the students and ensure complete reporting of the use of free time. In addition, the students were asked to rank the three of these activities that they considered most helpful to their professional development, with a ranking of 1 indicating most helpful. For each senior student, data were obtained on residency type and location, age, sex, performances on the National Board of Medical Ex-